7. The Theory of Probability

7.1 Introduction

In investigating weighing, and sorting and finding an efficient code, you may note a common thread.

In each case, and in many others that we will encounter later, we start with a (usually large) number of possible situations, and are interested in determining the true situation from among them. Thus we have devised schemes for doing so in the contexts mentioned.

Probability theory, which by the way was developed initially in the study of gambling schemes, is intended to provide a framework for modeling problems of this kind, with the aim of providing information about what we can "expect" to happen in each context.

This is done by defining a **Sample Space** whose points or elements are the possible situations, each one of which is given a weight or probability according to the proportion of the time we expect it to be the true situation.

Then we determine what to "expect" as the answer to some question, by representing the answer as a function defined on the points of the sample space, and averaging that function over them.

In most contexts, including gambling and all that we have considered so far, the potential situations can be represented by strings of binary bits (or ternary symbols, in the case of weighing). In the following discussion we assume that our sample space consists of points each of which is such a bit string.

The words used in probability theory are as follows:

7.2 Basic Definitions and Facts

An **event** corresponds to a subset of the sample space; here are some examples:

1. the second bit of our bit string is a 1.
2. half the bits are 1 (obviously only possible when the string is finite and has an even length)
3. any other property of points in the sample space.

The subset corresponding to an event E consists of the points of the sample space which have the property that defines E.

A **random variable** is the name we give to a function defined on the points of the sample space. Here are some examples:

1. the number of 1's in the bit string
2. the position of the first 1 in the string
3. how much you would win in your betting scheme if the given bit string came true
4. any other (usually numeric) function defined on strings

A particularly useful kind of random variable is **an indicator random variable for an event E**. It is **1 on strings in the event set, and 0 elsewhere.** We can call it I(E).

As a sort of converse of this definition, given any random variable, say f, we can, for any of its values, x, **define E(f=x) to be the event that occurs when f(p)=x holds (and only then).**

We can then write

$$f = \Sigma_{\textbf{(over all x)}} \quad \textbf{x*I(E(f=x))}$$

Each point q of the sample space, that is, each possible bit string q, is given a weight, called its **probability**, which is usually written as p(q). **These probabilities are required to be non-negative numbers whose sum, over the entire sample space, is 1.**

A **uniform sample space**
is one in which all points have the same weight, which is therefore the reciprocal, 1/N, of the number, N, of points in the space.

Each **event E** is given a **probability, p(E),** which is **the sum, over all strings in its set, of their individual probabilities, p(q).**

The **expectation of a random variable f** is the **sum over the points q of the sample space of f(q)*p(q).**

This e**xpectation** is often denoted as **f with a bar over it**, or as **⟨f⟩.** It is often called **the mean** or **average** of the random variable f.

The **variance** of a random variable f is **the difference between the expectation of its square and the square of its expectation**, or $\langle f^2 \rangle - \langle f \rangle^2$. It is often denoted by and called **sigma squared of f**, or $\boldsymbol{\sigma^2(f)}$.

Another expression for the variance is $\langle (f - \langle f \rangle)^2 \rangle$. **⟨f⟩ is** often called **the mean value of f, f-⟨f⟩** is called **the deviation of f from its mean**, so that the variance can be called the "**mean square deviation of f from its mean**". Sigma of f is then "**the root mean square deviation of f**".

The expectation of f is by definition linear in f. This means that the expectation of the sum of af and bg for numbers a and b and random variables f and g obeys ⟨af+bg⟩ = a⟨f⟩ + b⟨g⟩.

Notice that the variance of a function f is quadratic in f, and we have
$$\boldsymbol{\sigma^2(af) = a^2\sigma^2(f),}$$

when a is a constant. On the other hand the variance of a sum of two random variables is often much smaller than its quadratic nature would suggest.

Notice that the expectation of an indicator random variable (whose values are all 0 or 1), is the probability of its event.

Recall that an event E corresponds to a subset of our initial sample space, S. We define the **conditional probability of any event Q given E** to be the **probability of Q in a new sample space defined to be the points on which E happens.**

This probability is usually written as p(Q|E).

The probability of Q happening in general is the sum over all points p in S of p(p)*I(Q) divided by the sum over all p of p(p) which latter is 1.

The conditional probability p(Q|E) has numerator given by the probability that both Q and E occur, and denominator given by the probability that E occurs.

Thus the indicator random variable for the event Q given E is the product I(E)*I(Q), which is the indicator of the event that both Q and E take place. The probability of the event is the sum of this random variable multiplied by p(p) over all points p of E (or all points of S, same thing) divided by the sum of I(E)*p(p) over all p in E.

The denominator is required here because we have reduced the entire sample space S to the points in E. The sum of p(p) for our old probabilities over our new sample space will not usually be 1. To get a probability distribution that sums to 1 while maintaining the relative probabilities of points in E, we must take the new probabilities of each point p in E to be p(p)/(SUM(over p in E of)p(p)).

Perhaps the simplest example of this is as follows: Suppose we toss two fair coins and want the probability that both are heads. This is one fourth. If we toss one and it is a head, then the probability that both are heads, given this information becomes ½. This is what the remarks of the preceding paragraph imply.

These remarks can be written in symbols as

$$p(E|Q) = p(E \ \&Q)/p(Q).$$

There are analogous statements about random variables. You can speak of the probabilities that a random variable f takes on each of its variables, given that some event takes place.

On occasion you may want to express the probability of E given Q in terms of the probability of Q given E.
From the last relation we can write:

$$P(E|Q) = p(E\&Q)/p(Q) = (p(Q|E)/p(Q)) *p(E) .$$

In words, this says that the probability of E given Q differs from the probability of E at the beginning, by a factor that is the ratio of the probability of Q given E to the probability of Q at the beginning.

Two events are said to be **mutually exclusive** when both together happen on no point of our samples space S.
We can deduce that the probability that either of two **mutually exclusive** events occur is the sum of the probabilities that each occur.

$$\text{If } I(E)*I(Q)= 0 \text{ then } p(Q \text{ or } E) = p(Q) + p(E).$$

More generally when Q and E can both occur simultaneously, the points of the sample space on which both occur are counted twice on the right hand side here.

The general statement is therefore:

$$P(Q \text{ or } E) = p(Q) + p(E) - p(Q\&E).$$

We can write out a similar statement when we have three or more events.
With three events, the probability that any of them occur is the sum of the probabilities that each one occurs less the sum of the probabilities that two of them occur, and that will correctly account for all cases in which at most two of them occur. However when all three occur we should have a contribution but have added three and subtracted three contribution; so we must add it back again. We get

$$P(Q_1 \text{ or } Q_2 \text{ or } Q_3) = \Sigma_{(over\ j)}\ p(Q_j) - \Sigma_{(over\ j<k)}\ p(Q_j\&Q_k) + p(Q_1\&Q_2\&Q_3).$$

Exercise 7.1: What is the analogous statement for 4 events? For n events?

    <u>**Two events A and B are said to be independent**</u>, if the probability of A given B, $p(A|B)$, is exactly the same as the initial probability, $p(A)$.

    We may deduce that
the probability both A and B happening when A and B are independent is the product of their individual probabilities. For, the statement of independence is

$$P(A) = p(A/B) = p(A\&B)\ /\ p(B).$$

    Independence can also be defined for random variables. **Two random variables f and g are said to be independent when learning the value of f does not change the probabilities that g take on any of its values.**

    This means $p(f=a) = p(f=a|g=b) = p(f=a\ \&g=b)/p(g=b)$, or
$p(f=a\&g=b) = p(f=a)*p(g=b)$, which implies that for independent random variables we have

$$I(f=a,g=b) = I(f=a)*I(g=b).$$

    If we multiply both sides here by a*b and sum over all a and b, we obtain **the condition that random variables f and g are independent means**

$$\langle f*g \rangle = \langle f \rangle * \langle g \rangle,$$

    the expectation of their product is the product of their expectations.

We now ask the question: **what is the variance of the sum of two random variables, f and g?**

    From its definition we have

$$Var(f+g) = \langle (f+g)^2 \rangle - (\langle f+g \rangle)^2,$$

and using the linearity of the expectation and separating the f g and cross terms we get

$$Var(f+g) = \langle f^2 \rangle - \langle f \rangle^2 + \langle g^2 \rangle - \langle g \rangle^2 + 2(\langle f*g \rangle - \langle f \rangle * \langle g \rangle).$$

$$= \ Var(f) + Var(g) + 2(\langle f*g \rangle - \langle f \rangle * \langle g \rangle).$$

The **covariance of f and g,** Cov(f,g), is defined to be
$$\langle f^*g \rangle - \langle f \rangle^* \langle g \rangle,$$

so that we get

$$\textbf{Var } (f+g) = \textbf{Var}(f) + \textbf{Var}(g) + 2\textbf{ Cov}(f,g).$$

Notice that the covariance of independent variables is 0. And so, **when f and g are independent,** we get

$$\textbf{Var}(f+g) = \textbf{Var}(f) + \textbf{Var}(g).$$

This innocent looking statement is actually very curious and important. We have here a linear type property of a quadratic function of f and g. It is the basis of the results we will need, which concern sums and averages of many independent random variables.

7.3 Comments

A superficial look at the preceding suggests that we have made lots of definitions but gotten essentially nowhere with them.

Actually we have gotten very far, as we shall soon see.

But you must realize that in fields like probabilities words are introduced that suggest meanings, which in fact they may not actually have. For example, we have introduced the concept of the *expectation value*
of a random variable. Despite the name, it does not necessarily mean that it is what we expect the variable to be, or that it has any meaning at all.

The expectation of a random variable is its average over your sample space with each point q weighted by its specified probability, $p(q)$. This may or may not mean anything at all. For example, an indicator variable for an event always takes values 0 and 1. Yet its expectation value is the probability of the event, which is the sum over the points of S of $p(q)^*I(E)$.

On the other hand, if a the variance of a random variable f is sufficiently small, the random variable will be near the expected value at most points of the sample space.

If the actual case we are interested in is typical of the point in the sample space, the expected value of f does tell us what to expect from f, when the variance of f is small enough.

In order to draw conclusions from probability theory, we will need one more tool: which tells us how we can draw conclusions about the meaning of the expectation of f from information about its variance.

Our goal here is to be able to analyze the behavior of situations that can be described as long sequences of independent (or mostly independent) random variables.

In particular we will be interested in the situation in which we have a long string of bits which contain a relatively small number of corrupted bits, and also of properties of randomly chosen bit

strings, of a given length.

What we can say in such situations depend on what assumptions we can make about them.

If we are concerned with a sequence of length N of independent 0 1 valued random variables, each with the same probability of being 1 (these are called identically distributed independent Boolean random variables) then we can make very strong statements about their sum and their average. We can in that case determine the probability that their sum or average takes on any value.

As we relax the conditions here, so that the variables do not have to all be identical, and can take on other values, and can have limited dependence on one another, we can say somewhat less, but for a wide variety of circumstances we there is s similar conclusion, called the central limit theorem. It has many variations based on varying assumptions about the situation modeled, usually with roughly the same conclusion.

There is a weaker statement called the "law of large numbers" which is sufficient for our purposes here. It states that if you choose a value from a set of independent identically distributed random variables many times, the proportion of the time that you get any particular value will be close to the probability of that your random variable takes on that value, almost all the time, as your number of picks increases. (Actually there are several laws of large numbers.)

This statement can serve as a justification for interpreting the probability of an event as the proportion of the time the event would occur, if you kept putting yourself in the situation modeled by your sample space, over and over again, independently.

The rest of the chapter consists of a description of our last tool, and then some of the consequences that can be deduced from all this, about sequences of random variables. We leave you with some problems that can be solved by clever use of the ideas we have discussed so far, namely:

1.
  the expectation of a sum of several random variables is always the sum of their expectations.
2.
  the expectation of a product of independent random variables is the product of their expectations,

7.4 The uses of the variance. Tchebychev's inequality

The definition of the variance of f is equivalent to its interpretation as the mean square of f's deviation from its mean.

However, it can be used to put an upper limit on the probability that a random variable takes on values with deviations greater than $x\sigma$ for any x greater than 1.

The simplest and worst such bound is called Tchebyshev's inequality.
It is an answer to the question:
how much of the probability of a variable can correspond to a deviation greater than $x\sigma$?
You can see right away that if you want to minimize the mean square deviation, and have a given amount, z, of deviation greater than $x\sigma$, you are best off having all the deviation that is less than $x\sigma$ have 0 deviation, and the rest have deviation just barely greater than $x\sigma$.

The mean square deviation in that case will then be $z(x\sigma)^2$ **which can be at most σ^2.**

We deduce from this statement that **the probability, z, that the deviation is greater than $x\sigma$ is at most x^{-2},** and this is **Tchebychev's inequality.**

This inequality tells us that

the probability that your random variable differs from its mean by more than $x\sigma$ is at most x^{-2}. **When σ is very small,** this statement implies that **your random variable is close to its mean over a portion of your sample space that has high probability.**

And that kind of statement, **when σ of your variable goes to 0 as the length of our bit string increases,** is called a **Law of Large Numbers.**

7.5 Some Conclusions.

Suppose first that
we have N independent variables each having values of 0 and 1 only, and each with a probability p of being 1.

The sum of these variables will be k when k of them are 1 and N-k are 0.

Every specific way that this happens has probability given by the product of the probabilities of each variable being what it is, and that will be $p^k(1-p)^{N-k}$ since the variables are all independent of one another.

There are C(N,k) different ways of this happening and these are mutually exclusive, so that their probabilities add, and **the probability that their sum is k is $p^k(1-p)^{N-k}C(N,k)$.**

Thus, in this case we deduce:

p(Sum=k) = p(Average=k/N) = is $p^k(1-p)^{N-k}C(N,k)$.

This is called a <u>binomial distribution</u> and its histogram goes up and comes down with a bell shape, for large N and fixed p.

We know from our earlier mutterings, that **the mean of this distribution is the sum of the means of its individual summands, which are p, so that the mean of the sum is pN, and the mean of their average is p itself.**

Furthermore, since these variables are assumed to be independent, the variance of their sum is the sum of their individual variances, p(1-p), or Np(1-p), **and the variance of their average is p(1-p)/N.**
(remember that the variance is quadratic, so that dividing the sum by N, which is the way we average variables, divides the variance by N^2)

This tells us, by Tchebychef's inequality, that the average of our N variables approaches p except on a set whose probability approaches 0, and this is the **"Weak Law of Large Numbers."**

This statement provides a meaning of sorts, to the probability p of a single variable. It is also exactly the statement we will need in the next section.

Similar results to all of those above hold when assumptions are weakened. Here all variables are assumed to be independent, and identically distributed.

Thus, as long as the variables are independent and each of their variances are small compared to the sum of the variances, it is possible to prove that the resulting probability for the average of the variables, is not unlike a binomial distribution, except it will have mean given by the average of the p's and variance given by the sum of the individual variances divided by N^2.

This holds true for sums of random variables which take on more values than 0 and 1, and even (with some changes) when there is a limited amount of dependence among the variables, as well.

Results of this kind are called central limit theorems.

When we add a sum of a large number of independent and identically distributed random variables, and the sum of their means stays finite, we get a special case of a binomial distribution, called a Poisson distribution.
Wh
en the mean approaches m, this distribution has the form

$$p(k) = (\exp(-m))\ m^k/k!.$$

Again, distributions can tend to this one with much weakened assumptions about them.

You should also realize that the facts about probability that we have noticed above, that the mean of a sum is the sum of the means in all cases, that the variance of the sum of independent variables is the sum of the variances, etc. are extremely useful tools for solving probabilistic questions, as you will see in the exercises below.

Exercises:
7.1 is buried above. Find it! Do it!
7.2 use a spreadsheet to plot (chart x y scatter) p(k/N) vs k for binomial distributions having p=1/2 and p=1/3 for values of N from 5 to however high you can, on one sheet for each p.
7.3 . Derive the formula for a Poisson distribution from the binomial distribution with p=m/N, for finite k as N gets large.
7.4 Given a class of N students with independently chosen birthdays, deduce the probability that no two have the same birthday. This can be done by deducing a formula for it and taking logs and expanding them to approximate the answer.
Another way to estimate the answer is by assuming you have a Poisson distribution of the number of pairs of students with the same birthday. How do these answers compare?
7.5 a. Suppose you pick 2 numbers between 0 and 1 at random (that is, the probability of picking a number in any interval of length d is 1/d) for each). What is the expected value of their difference? Hint: what is the answer if you pick 3 points at random on a unit circle?
b. Suppose you pick 2 distinct numbers at random between 1 and N. What is the expected value of their difference? Hint: compare with part a.
c. Suppose you pick k distinct numbers at random out of N? same question.
7.6 Suppose k persons get on an elevator in the basement of an office building one mornings and they get out each with equal probability of emerging at each floor there being m floors at which

they might get out, and suppose they get out independently. Find the expected number of floors at which the elevator does not stop, because nobody gets out.

Evaluate for m=11 and k=10 and vice versa.